Overview of BCHM3071

- **Gene Architecture and Regulatory Mechanisms** (10 Lectures)
  - Genes and Genomes
  - Chromosome Structure
  - Regulatory Elements in DNA
  - Regulatory Proteins
  - Mechanisms of Action of Regulatory Proteins
  - The Histone Code
  - Transcription Factors in Development
  - Epigenetics in Human Health and Disease
  - X-inactivation
  - Parental Imprinting

- **RNA in the Control of Information Flow** (4 Lectures)
  - Splicing Engines
  - RNA Stability
  - small RNAs 1: miRNA and CRISPR
  - small RNAs 2: siRNA and piRNA

- **Transcriptomics** (4 Lectures)
  - RNA Biology and the Complexity of Genomes
  - RNA Editing
  - RNA Switches (Riboswitches)
  - "Omics" Approaches (RNomics)

- **Maintenance of a Dynamic Genome** (6 Lectures)
  - Homologous Recombination
  - Molecular Machines in Recombination
  - Short-range Repair of DNA
  - Model Systems for DNA Repair
  - Mobile Elements
  - What happens when our DNA changes?

Lecture 1 (Genes and Genomes)

Revision of characteristics of eukaryotic genes; arrangement of genes in different organisms, size and complexity of genomes; regulatory regions; repeat elements, retrotransposons, repetitive DNA, centromeres, telomeres. Experimental evidence: sequencing projects, homology searching.

Gene

- General term describing any stretch of DNA that **encodes diffusible products**
  - Products include protein and RNA (rRNA, tRNA, miRNA etc)

Expressed Sequence Tags (ESTs)

- mRNA gives indication of the protein coding capacity
- ESTs → using mRNA sequences to ID no. of genes in a particular genome

The Human Genome

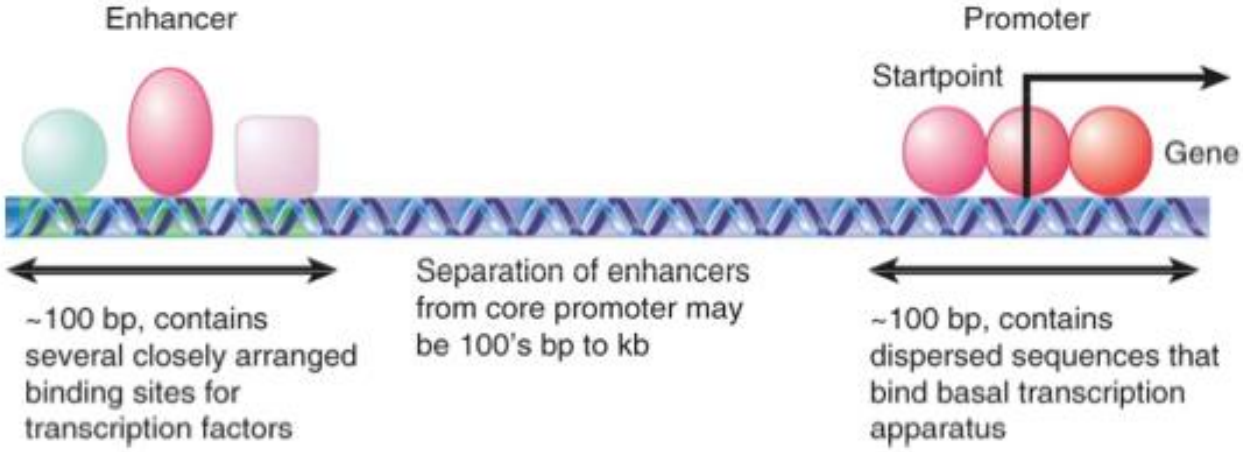
- Ave. gene size ~1000-2000 bases (*exonic sequence ONLY*)
- Approx. 20,000-30,000 genes
  - 1000 bases x 30,000 genes → **30Mb (30,000,000 bases)**
- Human haploid genome **~3000Mb**
- LOW GENE DENSITY → Exonic sequence only constitutes **1-2% of the genome**

What about the rest of the Human Genome?

- **Regulatory regions (Promoters, enhancers etc.)**
- Introns
- **Centromeres, telomeres, Origins of Replication**
- **Pseudogenes**
- **Short repeat sequences**
- **Long repetitive elements (LINEs, SINEs, transposons)**
- Other Intergenic DNA

Regulatory Regions

- Each gene has its own promoter(s) + enhancer(s) sequences

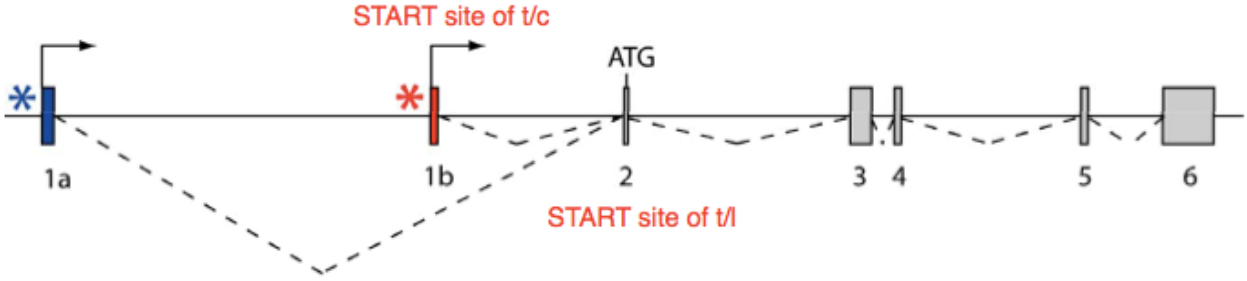


Promoters upstream of START sites of t/c (NOTE: START site for t/c is UPSTREAM from the START codon for t/l)

- ~100bp
- Binds basal t/c apparatus

Enhancers

- Contain binding sites for t/c factors → encourage the recruitment of DNA pol.
- Separated from promoters by ~100's bp to kb UPSTREAM
- Sometimes enhancers can be within introns (DOWNSTREAM of promoter)
  - Normally contained in the 1st intron



Alternative promoters/enhancers may be employed to achieve SPECIFIC expression patterns

- Create diversity and flexibility in the regulation of gene expression

Introns

- No. of introns >>> No. of exons
- Introns *may* contain regulatory regions (enhancers, alt. promoters)
  - Contribute to stability of the genome
- Allows for **EXON SPLCING** → give rise to protein isoforms
  - Nearly 80% of genes are subjected to alt. splicing → ↑ Protein coding capacity

Centromeres, telomeres, Origins of Replication

- Regions of DNA having ***chromosomal functions*** → maintain integrity of the chromosome
  - **Centromeres** → Chromosomal segregation during mitosis
    - Mitotic spindles attaches to centromeres
  - **Telomeres** → prevent chromosomal end shortening
  - **Origin of Replication** → docking site for DNA pol. during replication
- Characterised by short, tandem repeats
- Telomeres also take up space in our genome

Pseudogenes

Copies of functional genes → however ***altered so that they NO LONGER have function of parent gene***

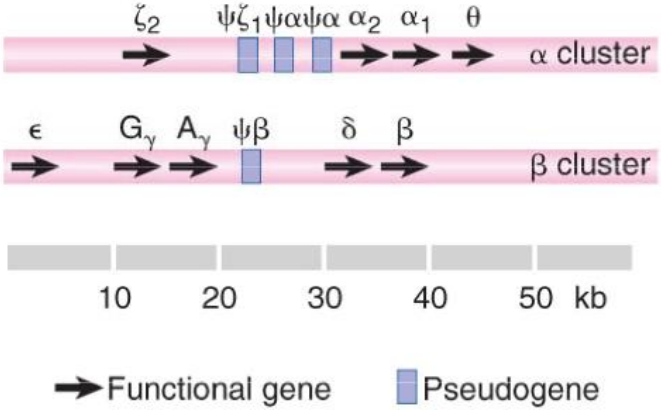
- Nonprocessed pseudogenes
- Processed pseudogenes

Nonprocessed pseudogenes

- **Gene duplication** followed by inactivating mutation OR incomplete duplication

For example, both the *alpha and beta clusters* of the **human globin locus:**

- Contain pseudogenes which are similar to functional genes
- However mutation renders them **non-functional**



For example, **β-globin gene (β)** [picture above], versus **β-globin pseudogene (ψβ)** → any changes in the features of the active gene will render it non-functional

- Promoter mutation → no longer t/c'ed
- Splicing junction lost → introns retained
- Point mutations
  - Nonsense → premature STOP codon
  - Missense → wrong AA coded

Processed pseudogenes

- Doesn't arise from direct duplication (unlike nonprocessed pseudogenes)
- Resultant from **reverse t/c of an mRNA + insertion** back into the genome
- The resultant processed pseudogene will be a continuous stretch of exons
  - Even though it has functional exons, the lack of introns and regulatory regions renders it non-functional
- We have ~200,000 pseudogenes in our genome

Short Repeat Sequences/Variable Number Tandem Repeats (VNTR)

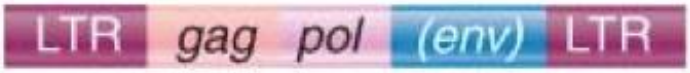
- Elements within the genome that are repeated many times over [take up space in the genome]
  - **Telomeres** (Satellite DNA)
  - **Microsatellites** = <10 bp repeating units → ↑ Mutation rate → High genetic diveristy
  - **Minisatellites** = ~10-100 bp repeating units
- UNSTABLE
- Any 2 individual will have DIFFERENT numbers of VNTR regions
  - Used in maternity tests

TRANSPOSABLE ELEMENTS (TE): Transposable Long Repetitive Elements

- Take up ~44% of the human genome, although very little remain active
- Class I TE (**Retrotransposons**) → Copied in 2 stages
  - With LTRs
  - Without LTRs
    - LINEs
    - SINEs
- Class II TE (**DNA transposon**) → Cut n Paste

Element	Organization	Length (Kb)	Human genome	
			Number	Fraction
Retrovirus/LTR retrotransposon	LTR gag pol (env) LTR	1–11	450,000	8%
LINES (autonomous), e.g., L1	ORF1 (pol) $(A)_n$	6–8	850,000	17%
SINES (nonautonomous), e.g., Alu	$(A)_n$	<0.3	1,500,000	15%
DNA transposon	Transposase	2–3	300,000	3%

Long Terminal Repeat (LTR) retrotransposons



- Code for **reverse transcriptase (RT)** and **integrase**
- Similar to retroviruses (but non-infectious)
- Ability to make DNA copies of themselves and integrate back into the genome → explains their abundance in the genome

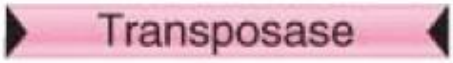
Non-LTR retrotransposons → **retroposons**

- Not flanked by LTRs at terminals

- **LINEs (Long-INterspersed Elements)** → encodes a reverse transcriptase
  - **Autonomous**

- **SINEs (Short-INterspersed Elements)** → DOES NOT encode their own reverse transcriptase
  - Relies on other mobile elements for transposition → **Nonautonomous**

DNA Transposons



- A sequence of DNA that can change its position within a genome = *"JUMPING GENES"*
  - Transposition usu. lead to duplication of the transposon
- DNA elements that encodes a ***transposase***
- ***Transposase*** → Cut n Paste mechanism to move the transposon and insert into a new site in the genome

TE → IMPORTANT for regulation of genome function and evolution

- In diseases → TE = Mutagen
  - Insertion into functional gene will affect/disable said gene

Lecture 2 (Chromosome Structure)

Euchromatin/heterochromatin, histones, structure of the nucleosome, structure of histones, modification of histones, DNA methylation and distribution of CpG

Euchromatin/heterochromatin

Chromatin = complex of DNA and proteins in the nucleus
Euchromatin = light staining → accessible DNA (Active genes)

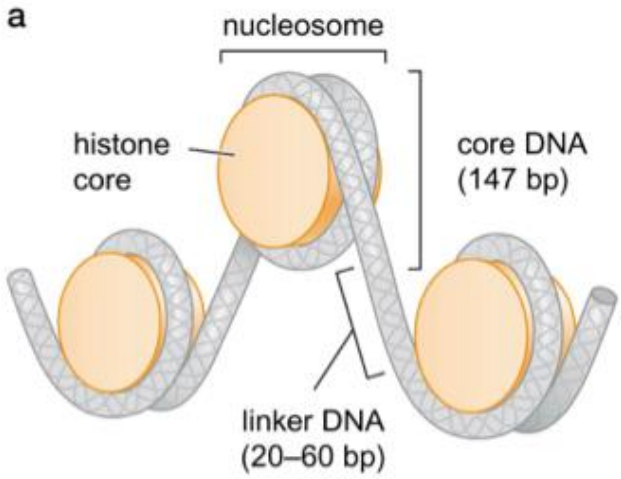
- Euchromatin however, doesn't always guarantee expression

Heterochromatin = dark staining → inaccessible DNA (Inactive genes)

- Stain more heavily due to tightly compacted DNA

Packaging

Chromatin = composed of **nucleosome** subunits
Nucleosome = DNA + histone core



Histone proteins

- Basic nature (+ve) due to lysine and arginine residues
- **Core proteins (Octamer)** → H2A, H2B, H3, H4
  - ~146 bp
- **Linker protein H1** → interact with DNA and the core octamer
  - 8 - 114 bp

Experiment showcasing DNA packaging

1. Chromatin extracted from cells
2. Digest chromatin with ***micrococcal nuclease (MNase)***
   a. ***MNase*** = non-specific cutting/random
3. Separation by gel electrophoresis

→ Found series of DNA bands, increments of ~200 bp

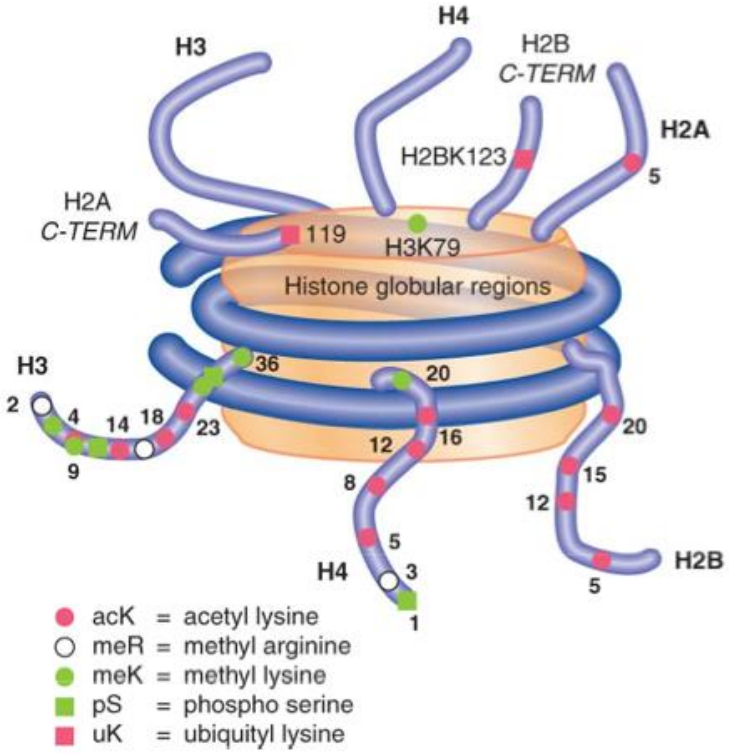
- Shows that nucleosomal core DNA is protected from MNase
- However, if you let MNase to completion → every linker DNA cleaved → only ONE band @ ~205 bp

Under e- microscopy → nucleosomes appear as 10nm fibres

- In heterochromatin → nucleosome coil to form 30nm double solenoid (tight level of packaging) → further compaction

What determines the packaging of DNA and thus the accessibility of the DNA?

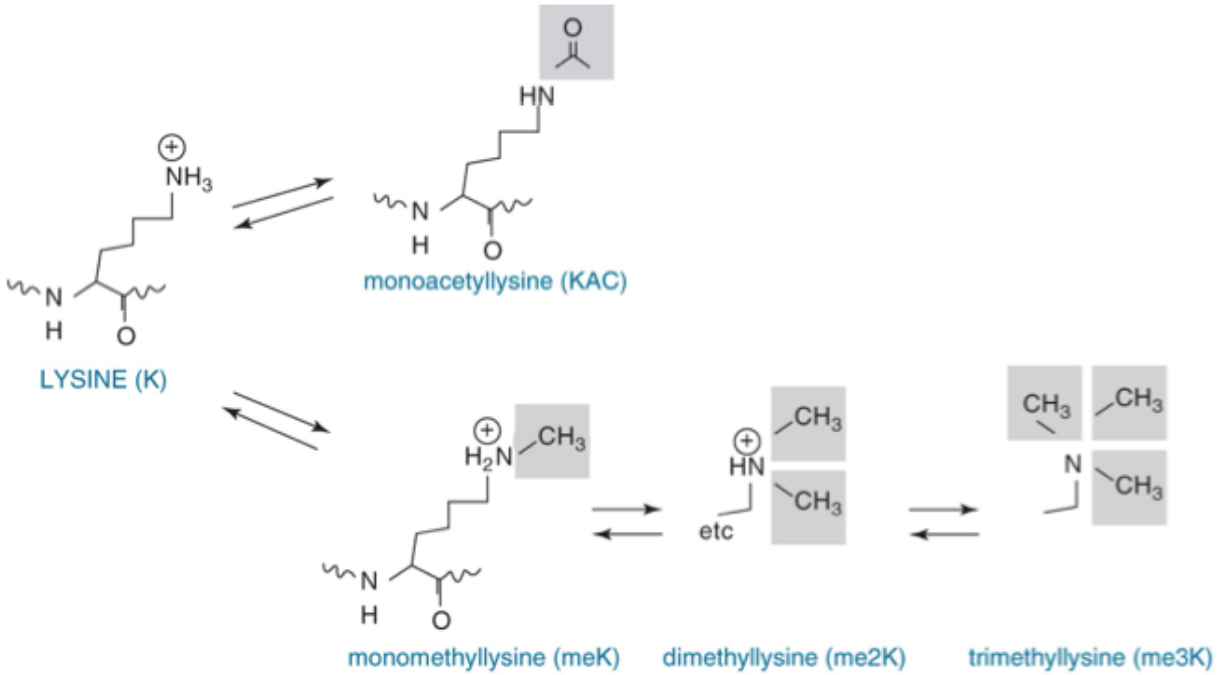
HISTONE TAILS



- Protrude from histone octamer
- Flexible in solution

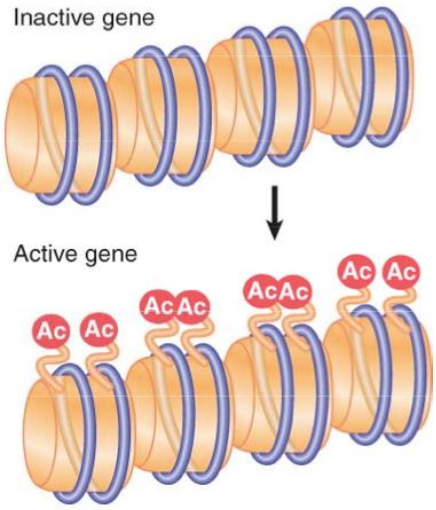
AA in histone tails are subjected to **post-translational modification** → contribute to packaging

- Methylation, acetylation, phosphorylation, ubiquitylation etc.
- Hundreds of combinations of modification → **Histone code** to packaging
    - Methylation alone can be meK, me2K, me3K (using lysine, K, as an example) → give rise to vast numbers of possible combination of the histone code for packaging



Acetylation of lysines in histones

- **Acetylated histones** = euchromatin/active genes
- Deacetylated histones = heterochromatin/inactive genes
    - Tightly packed → less likely to be expressed



Traditionally thought that acetylation of lysine was to neutralise the +ve charge on the lysine

- Now postulated that acetylation serves as a "signal" for packaging/un-packaging

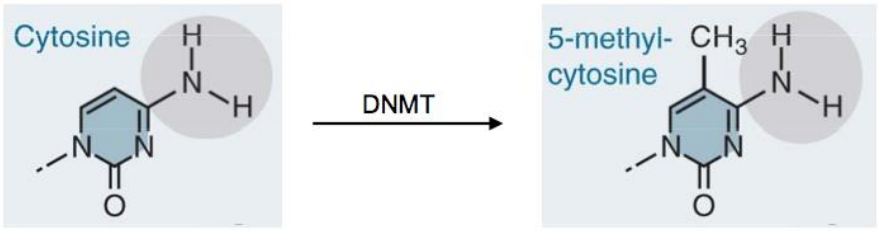
CpG dinucleotide and CpG Islands

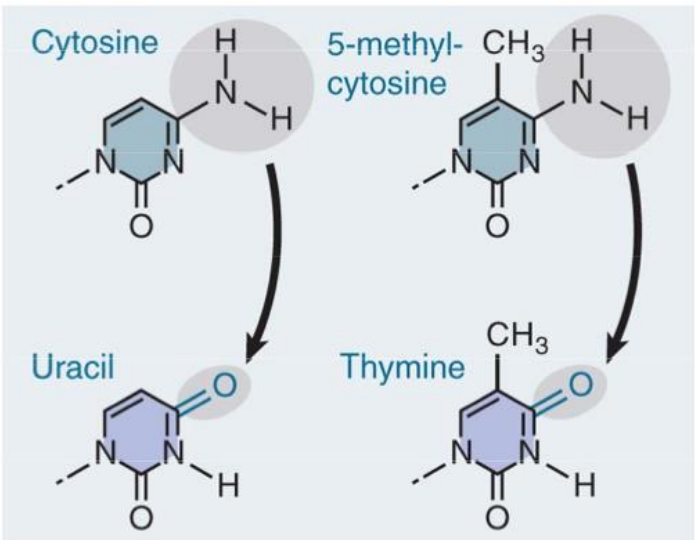
CpG dinucleotide

- Regions of DNA when C is followed by G in the linear sequence of bases along 5' → 3'
    - *Don't confuse it with CG base pairing*
- Short hand for *5' --- C --- phosphate --- G --- 3'*

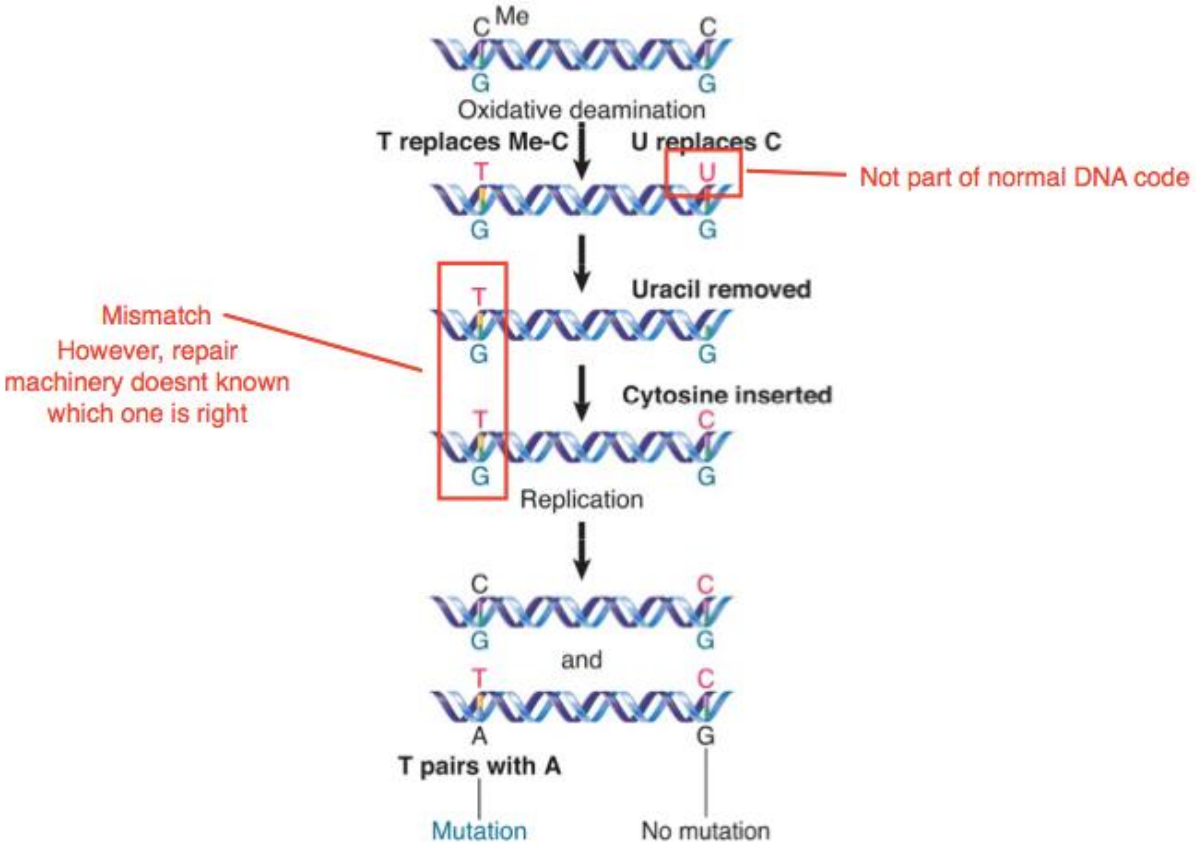
Subjected to modification

- Methylation of C by **DNA methyltransferase (DNMT)**
- DNMT exclusively methylate CpG





1. Cytosine can be spontaneously deaminated (oxidative deamination) to **Uracil**
    a. Uracil recognised as foreign (since it is RNA exclusive) → replaced by cytosine
2. $5^{me}C$ deaminated to **Thymine** (Lead to an *mismatched T-G base pairing*)
    a. Mismatch → repair machinery replace T or G (since it doesn't know which one is "right")
        i. If T replace → no mutation
        ii. If G replaced → Thymine persist → T-A base-pairing instead of C-G base pairing → **MUTATION**

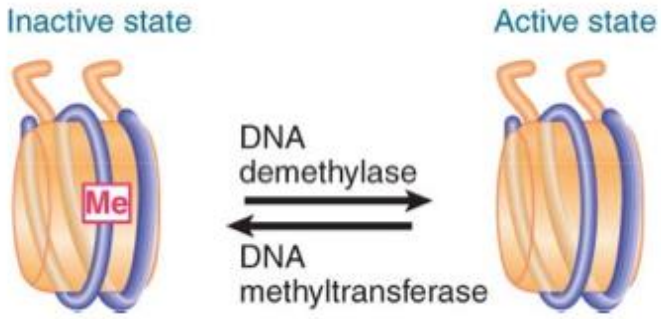


Thus methylation of cytosine in CpG sites = <u>lead to mutation</u>

- Scarcity of CpG dinucleotide in the genome due to the issue of spontaneous oxidative deamination
- Exception of CpG islands surrounding promoters

<u>CpG methylation and gene expression</u>

- Methylated CpG → Inactive state
- Demethylated/Unmethylated CpG → Active state



CpG Islands

- Regions of genome with a high freq. of ***UNMETHYLATED*** CpG dinucleotides
- Often in **promoter regions**, in particular housekeeping genes
- *Thought that CpG island is permissive to constitutive expression*
- CpG islands around active genes act as signal to RNA pol. indicating the start of a gene

For example, γ-globin vs APRT demonstrate CpG islands' permissive nature in housekeeping gene expression

γ-globin	APRT
• **Tissue-specific gene** • Low freq. of CpG dinucleotide	• **Housekeeping gene** (Constitutively expressed) • CpG island in promoter region

Protection of CpG islands from methylation

- Not 100% sure of the mechanism
- Mediated by t/c factors + protein cofactors
    - I.e. Sp1 binding to CpG

ADDITIONAL INFO

Promoter CpG hypermethylation in cancer

- *Suggests that tumour suppressor genes in cancer have lost expression due to CpG island methylation (inactivation)*
- *Hypomethylation → lead to overexpression??*